Protocol of data abstraction: occurrence and prevalence data on African animal trypanosomiasis

The protocol describes how data on the occurrence and prevalence of African animal trypanosomiasis (AAT) are assembled into a geospatial database within the framework of the "the Atlas of tsetse and AAT" initiative. The protocol draws on previous work carried out for other vector-borne diseases such as malaria [1] and for sleeping sickness, the human form of African trypanosomosis [2].
Data are extracted from publications stored in a central data repository (PDF format). Extracted data are assembled in a geographic database which includes three tables: one for data sources ('Sources'), one for geographical data ('Geo_data'), and one for data on AAT ('Epi_data'). Information items included in each of these tables are described below.

Data sources

The table 'Sources' provides details on the publications used to extract data on AAT distribution. The table includes the following fields:

1. SOURCE_ID. A unique numeric identifier of the source examined.
2. INITIALS. Initials of the first author. Where there is more than one initial, they are separated with a space, *e.g.* V A.
3. AUTHOR. Surname of the first author of the publication only.
4. ALL_AUTHORS. Full list of authors of the paper.
5. TITLE. Title of the paper.
6. YEAR. Year of publication of the paper.
7. COUNTRY: Country or countries for which the paper provides spatially referenced data on AAT.
8. AAT_DATA. 'Yes' identifies documents containing spatially referenced data on AAT occurrence and/or prevalence.
9. TSETSE_DATA. 'Yes' identifies documents containing spatially referenced data on tsetse absence presence and/or abundance (this flag is used to assist the development of the tsetse component of the Atlas)
10. TSETSE_INFECTION_DATA. 'Yes' identifies documents containing spatially referenced data on tsetse infection.
11. FILE_NAME: Standardized naming for PDF files, including author(s) and year of publication (e.g. Beadell_et_al_2010.pdf).
12. JOURNAL: name of the journal where the paper was published
13. PUBLISHER: name of the publisher of the journal
14. ACCESS_DATE: date when the paper was obtained/downloaded/provided from its source.
15. OPEN_ACCESS: 'Yes' identifies 'open access' sources.
16. EXTERNAL_LINK: Uniform resource locator (URL) for the source, if available. If the Document Object Identifier (DOI) is available, it is used preferentially to generate a permanent URL (e.g. http://dx.doi.org/10.1371/journal.pntd.0000636)
17. NOTES: the field contains all relevant comments concerning the source and its relevance for mapping AAT.

18. QUESTIONS: questions and requests for clarifications to the authors of the, which have the potential to contribute to the future refinement and improvement of the database

All sources are also separately stored in EndNote format, which enables additional information on the papers to be stored, as well as efficient extraction and formatting of the references.

Geographical data

The table 'Geo_data' contains detailed information on the geographical scope of study and it includes geo-positioning information. We define as 'site' the geographic location were the study was carried out, and it constitutes the basic unit for mapping. If more than one site is included in the study/paper, a separate entry for each location is recorded. All sites are represented as points in the Atlas of AAT, even when they refer to administrative areas such as districts or provinces. In the latter cases, the point normally refers to the centroid of the administrative area.
The following fields are recorded in the table 'Geo_data':

1. LOCATION_ID. A unique numeric identifier of the site.
2. SOURCE_ID. The numeric identifier of the source containing the site (it is extracted from the corresponding field in the table 'Sources').
3. COUNTRY. Country where the site is located.
4. LOCATION_NAME. Name of the site where the study was conducted. As a rule, the name as reported in the paper is recorded. Alternative spellings may be recorded in the field 'SITE_NOTES' (see below).
5. ADMIN1. Name of the first subnational administrative unit where the site is located (as reported in the paper).
6. ADMIN2. Name of the second subnational administrative unit where the site is located (as reported in the paper).
7. ADMIN3_PAPER. Name of the third subnational administrative unit where the site is located (as reported in the paper).
8. ADMIN2_ID: unique code for the corresponding admin2 as provided by the Global Administrative Unit Layers (GAUL) 2009 – reference year: 2008). The code is available from the field 'ADM2_CODE'.
9. LAT. Latitude of the study site in decimal degrees (Datum: WGS84).
10. LONG. Longitude of the study site in decimal degrees (Datum: WGS84). LAT LONG coordinates are specific to the site listed in LOCATION_NAME. More detailed information explaining how the coordinates were found and checked may be available in the field 'LOCATION_NOTES'.
11. GEO_SOURCE. Source of geo-positioning, which can be the paper itself or one of the many available gazetteers.
12. LOCATION_NOTES. Includes details on any decisions made while geo-referencing. This should help a third person understand the process and enable them to repeat the choices made.
13. AREA: Surface area of the survey (area of applicability for the recorded AAT data)
14. AREA_TYPE. Describes the size of the area of applicability for the recorded AAT data. This measure is not often explicitly detailed in the source and may need to be

inferred from GIS information as well as from context. Six categories are contemplated:

- ≤ 10 km^2;
- >10 and ≤ 25 km^2,
- >25 and ≤ 100 km^2
- >100 km2 and ≤ 500 km^2
- >500 km^2 and ≤ 1,000 km^2)
- > 1,000 km^2 and ≤ 10,000 km^2.

Data on African animal trypanosomosis

Data on the occurrence and prevalence of AAT are recorded in the table 'Epi_data'. The table includes:

1. SURVEY_ID: A unique numeric identifier of the survey. For different reasons, a single study/paper may include a number of separate surveys, which will result in different values for the SURVEY_ID. For example, if AAT is investigated separately for different animal species (e.g. cattle, goats, sheep, etc.), different SURVEY_ID will identify the data for the different species. Also, if more than one diagnostic technique is used in the same study and if this entails different estimates of AAT prevalence, different rows/SURVEY_IDs will be used to record the different estimates.
2. LOCATION_ID: The numeric identifier of the site where the survey was carried out (it is extracted from the corresponding field in the table 'Geo_data').
3. SOURCE_ID: The numeric identifier of the source containing the data recorded for the present survey (it is extracted from the corresponding field in the table 'Sources').
4. MONTH_ST. Starting month of the survey.
5. YEAR_ST. Starting year of the survey.
6. MONTH_EN. Ending month of the survey.
7. YEAR_EN. Ending year of the survey.
8. SAMPLE_SIZE: number of animals sampled.
9. SPECIES_AN: species of animal (e.g. Cattle, goats, sheep, pigs, etc...).
10. BREED_AN: animal breed. Because of the difficulties of harmonizing the descriptions and denominations of animal breeds, the sections of the paper referring to breeds are reported verbatim in this field.
11. AGE_AN: age of animals. Because of the difficulties of harmonizing the descriptions of animal ages, the sections of the paper referring to ages are reported verbatim in this field.
12. SEX_AN: sex of animals.
13. HUSB_AN: Animal husbandry system. Because of the difficulties of harmonizing the descriptions of husbandry system, the sections of the paper referring to husbandry systems are reported verbatim in this field. Attention is given to whether animals are kept in a prevalently sedentary system, or semi-pastoral/transhumant one, or pastoral. More in general, all aspects that have implications in terms of exposure to infection are noted, including grazing versus zero-grazing, extensive versus intensive, commercial versus small-hold enterprises, etc.

Infections with individual trypanosome species/subspecies/subgroups

14. Tv: Number of animals positive to the test for *Trypanosoma vivax*.
15. Tc: Number of animals positive to the test for *T. congolense*.

    a. Tcs: Number of animals positive to the test for *T. congolense* (savannah subgroup).
    b. Tcf: Number of animals positive to the test for *T. congolense* (forest/riverine subgroup).
    c. Tck: Number of animals positive to the test for *T. congolense* (Kenya coast/kilifi subgroup).
    d. Tct: Number of animals positive to the test for *T. congolense* (Tsavo subgroup). Although presently recognized as a strain of *T. simiae* [3-4], it is still recorded in the database as *T. congolense* Tsavo if so named in the data source.

16. Tb: Number of animals positive to the test for *T. brucei* s.l.
    a. Tbb: Number of animals positive to the test for *T. b. brucei.*
    b. Tbr: Number of animals positive to the test for *T. b. rhodesiense.*
    c. Tbg: Number of animals positive to the test for *T. b. gambiense.*
17. Tsi: Number of animals positive to the test for *T. simiae*.
    a. Tst: Number of animals positive to the test for *T. simiae* (Tsavo strain), initially classified as another type of *T. congolense* [5].
18. Tsu: Number of animals positive to the test for *T. suis.*
19. T: Number of animals positive to the test for any of the five species of trypanosomes under study (*T. vivax , T. congolense, T. brucei, T. simiae* and *T. suis*).

Note that animals diagnosed with a mixed infection (infections with more than one species/sub-species/subgroup of trypanosomes) are included in the counts for the respective infections with individual species.

20. TPR [%]: Total AAT Prevalence (in percentage), including infections with any of the five species of trypanosomes under study (*T. vivax , T. congolense, T. brucei, T. simiae* and *T. suis*).

Separate values for the prevalence of individual species/subspecies/subgroups of trypanosomes are also included in the database, but they are not listed in this document.

21. AAT_PRESENCE: 'Yes' if the African animal trypanosomosis is present, "No" if the disease is not present. This field is particularly useful when a paper only reported the absence or presence of AAT, but it reported neither the number of infections, not the prevalence.
22. DIAGNOSTIC: Diagnostic method. Because of the difficulties of harmonizing the descriptions and denominations of diagnostic techniques, the related sections of the paper are reported verbatim in this field.
23. PCV: Average Packed-Cell-Volume for the survey herd.
24. PVC_POSITIVE: Average Packed-Cell-Volume for the animals positive to the test for trypanosomosis.
25. PCV_NEGATIVE: Average Packed-Cell-Volume for the animals negative to the test for trypanosomosis.
26. TSETSE_INTERVENTIONS: it reports what interventions against tsetse were ongoing in the study area at the time of the survey, or in the recent past prior to the survey.

27. CHEMOTHERAPY: Includes information related to the use of therapeutic and/or prophylactic antitrypanosomal drugs.
28. SAMPLING_STRATEGY: it describes whether a random approach was used (e.g. in a study designed to assess the general epidemiological situation in an area), or a purposeful one (e.g. where specific villages/areas/herds/animals are investigated because of some peculiar features of theirs).
29. LONGITUDINAL: it describes whether data were extracted from a longitudinal study.
30. NOTES: Includes all important additional information as reported in the source paper.

References

1. Guerra C, Hay S, Lucioparedes L, Gikandi P, Tatem A, Noor A, Snow R: **Assembling a global database of malaria parasite prevalence for the Malaria Atlas Project.** *Malar J* 2007, **6:**17.
2. Cecchi G, Paone M, Franco JR, Fèvre E, Diarra A, Ruiz J, Mattioli R, Simarro PP: **Towards the Atlas of human African trypanosomiasis.** *Int J Health Geogr* 2009, **8:**15.
3. Stevens JR, Noyes HA, Dover GA, Gibson WC: **The ancient and divergent origins of the human pathogenic trypanosomes, *Trypanosoma brucei* and *T. cruzi*.** *Parasitology* 1999, **118:**107-116.
4. Gibson WC, Stevens JR, Mwendia CM, Ngotho JN, Ndung'u JM: **Unravelling the phylogenetic relationships of African trypanosomes of suids.** *Parasitology* 2001, **122:**625-631.
5. Majiwa PA, Maina M, Waitumbi JN, Mihok S, Zweygarth E: ***Trypanosoma (Nannomonas) congolense*: molecular characterization of a new genotype from Tsavo, Kenya.** *Parasitology* 1993, **106:**151-162.